Qiong Hu (405065032)
Zihao Zou (005349580)
Gaofang Sun (104853165)

Project 1 Report

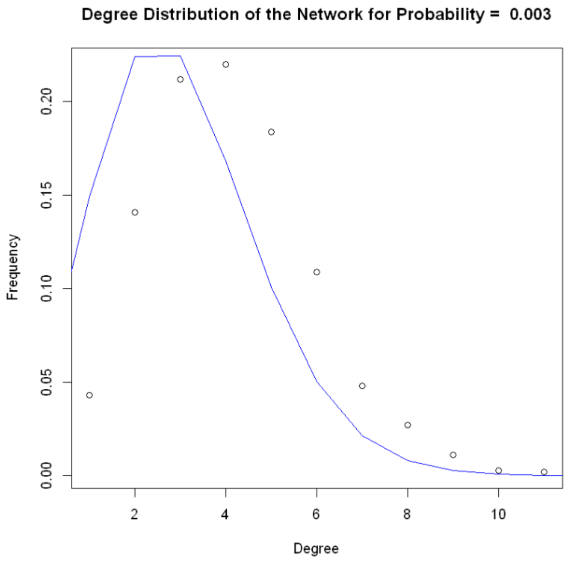
Random Graphs and Random Walks

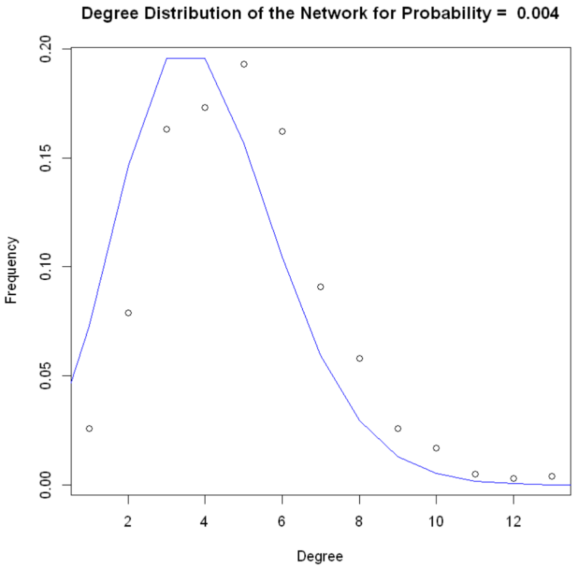
ECE 232E
Large Scale Networks
April 19, 2020

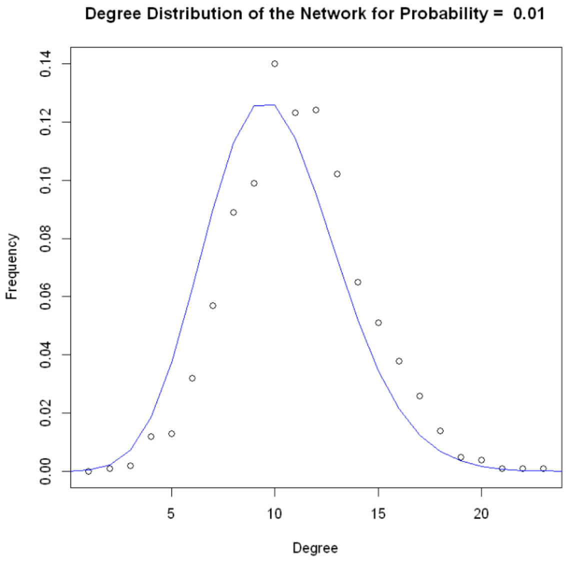
1 Generating Random Networks

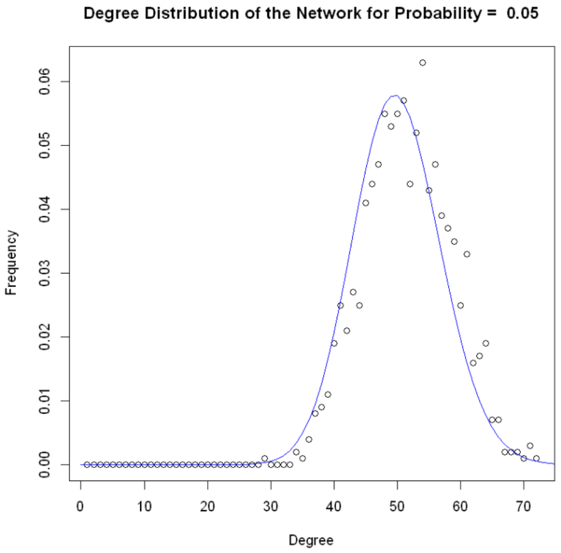
1. Create random networks using Erdös-Rényi (ER) model

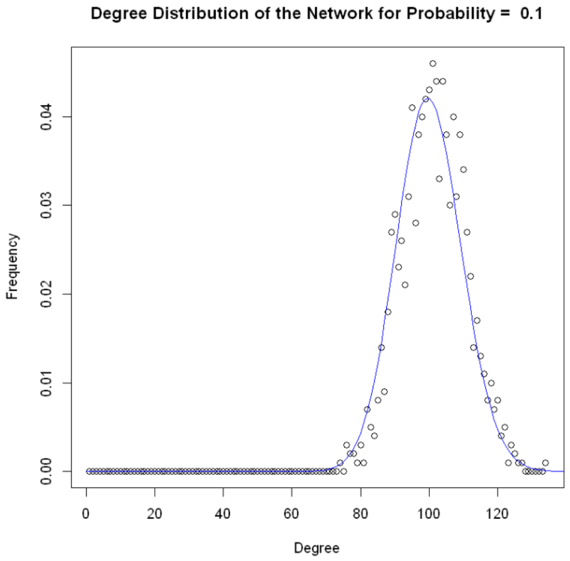
   (a) We create undirected random networks with n = 1000 nodes, and find the probability p for drawing an edge between two arbitrary vertices 0.003, 0.004, 0.01, 0.05, and 0.1 and plot the distributions. The blue line in each graph shows binomial distribution.


Degree Distribution of the Network for Probability = 0.003


Degree Distribution of the Network for Probability = 0.004


Degree Distribution of the Network for Probability = 0.01

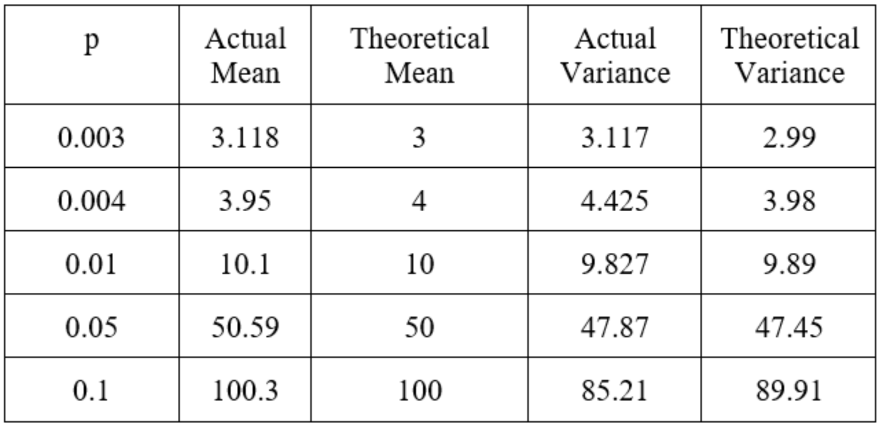

Degree Distribution of the Network for Probability = 0.05



From observation, we believe that ER network follows binomial distribution. The edges are independently with same probability p to exist. The total probability of drawing a graph with m edges is:

$$\Pr(m) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m}$$

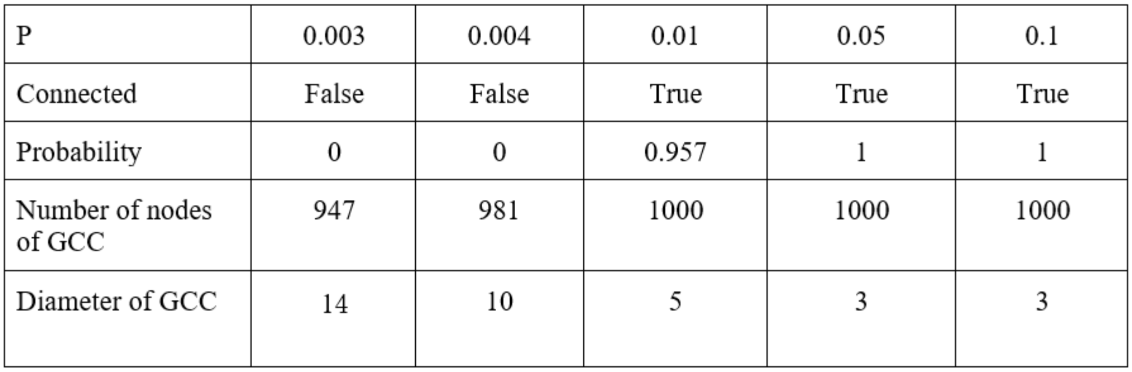
For the standard binomial distribution B(n, p), $\mathbb{E} = np$, and Var=$np(1-p)$.



p	Actual Mean	Theoretical Mean	Actual Variance	Theoretical Variance
0.003	3.118	3	3.117	2.99
0.004	3.95	4	4.425	3.98
0.01	10.1	10	9.827	9.89
0.05	50.59	50	47.87	47.45
0.1	100.3	100	85.21	89.91

As we can see from the table above, the mean degree distribution and variance of the generated network agrees well with the theoretical value. And for larger p, actual mean and variance are even closer to theoretical values.

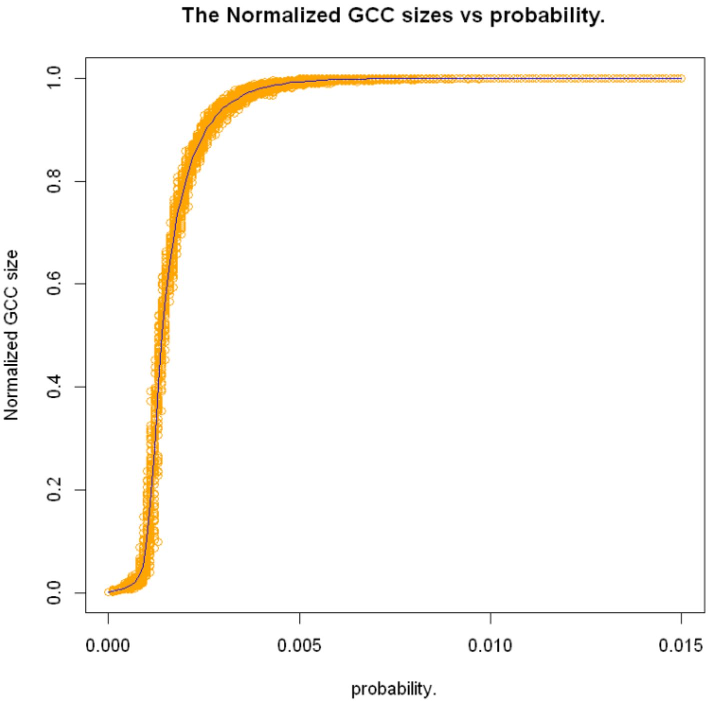
(b) For each p, we randomly generate network 1000 times to determine the probability of connection. The probability that a generated network is connected and the diameter of the GCC are shown below:



P	0.003	0.004	0.01	0.05	0.1
Connected	False	False	True	True	True
Probability	0	0	0.957	1	1
Number of nodes of GCC	947	981	1000	1000	1000
Diameter of GCC	14	10	5	3	3

From the table above, we find that not all random realization of the ER network are connected. When $p = 0.003$ or 0.004, the ER networks are always not connected. As the p increases, the connectivity probability gets larger, the number of nodes in GCC becomes larger as well, and the diameter of GCC gets smaller.

(c) From the previous problem, diameter of GCC is a nonlinear function of p. To observe this phenomenon, we set $n = 1000$, and sweep over values of p from 0 to p_{max} (we determined p_{max} as 0.15), which makes the network almost surely connected and create 100 random networks for each p. Then we scatter plot the normalized GCC sizes vs p and plot a line of the average normalized GCC sizes for each p along with the scatter plot.



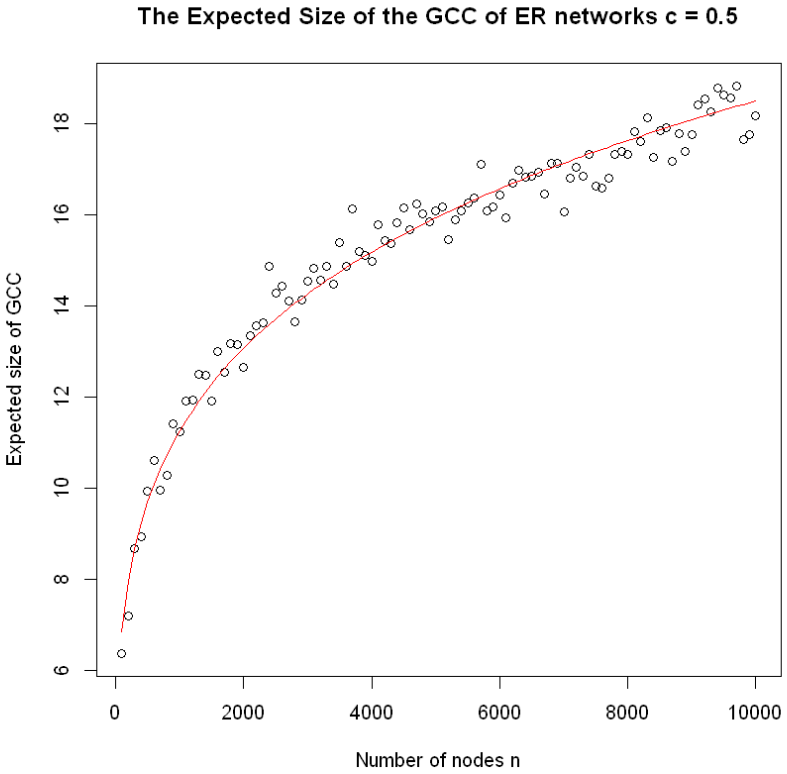
i. By observing the figure above, we found that the normalized GCC size increases sharply when probability p rising from 0.001 to 0.003. So we define "emergence" in this way:

when the size of the GCC grows the fastest, specifically, reaches the 50% of the total network size, we say it starts to emerge.

According to this definition and from the figure above, we empirically estimate that, when p reaches 0.001, the GCC starts to emerge. And this result matches the theoretical value $p = O(\frac{1}{n}) = 0.001$.

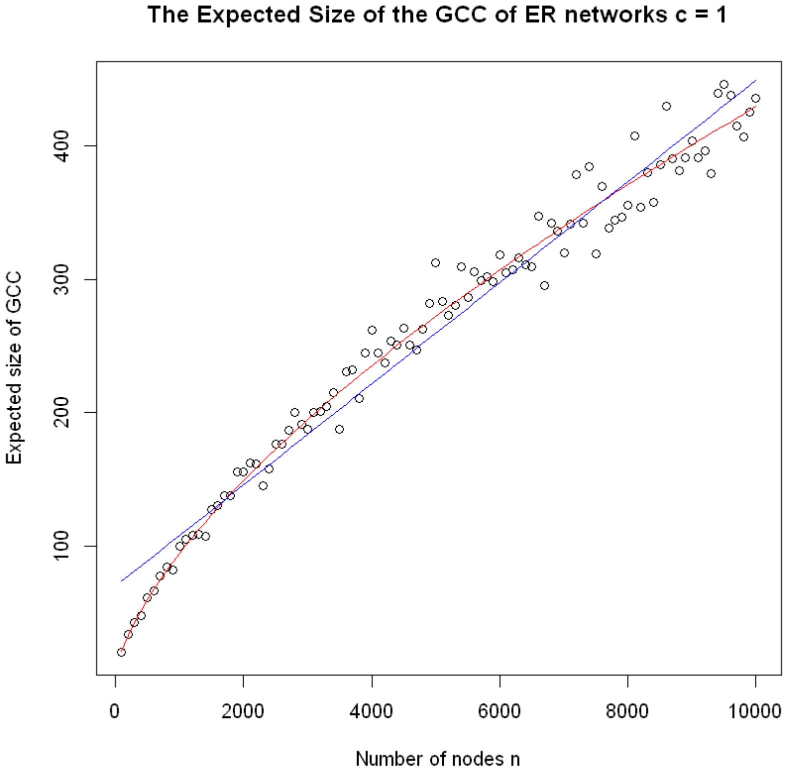
ii. At around $p = 0.004$, the GCC takes in over 99% of the nodes in almost every experiment.

(d) i. In this part, we define the average degree of nodes $c = n \times p = 0.5$ and sweep over the number of nodes, n, ranging from 100 to 10000, with a step of 100. Here, we plot every expected size of the GCC of ER networks with n nodes and edge-formation probabilities $p = \frac{c}{n}$ with 100 loops, and the trend is observed in the following figure.



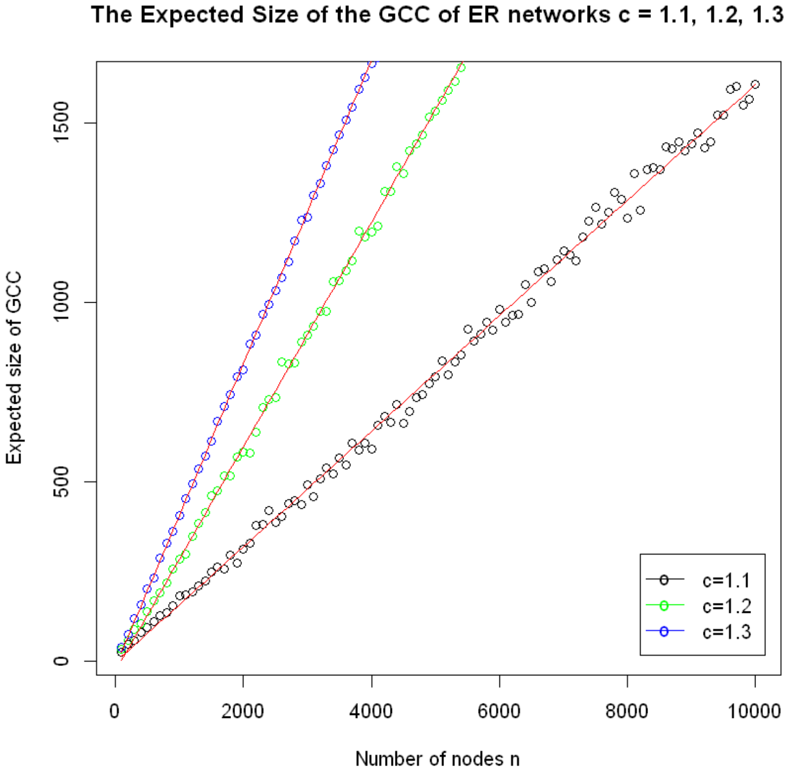
From the plot, we find that, in ER networks with fixed $n \times p = 0.5$, as n gets larger, the expected size of GCC gets larger, but the slope gets more smoother. The expected size of GCC is in the trend of $O(\ln(n))$, as seen in the plot where the red curve indicates the log fitting.

ii. We repeat the same for $c = 1$, as the plot and trend can be found below.



Similarly, we observe a positive increasing trend in the expected size of GCC with regard to n. However, the fitting curve here is somewhere between logarithm (shown in red) and linear (shown in blue).

iii. We repeat the same for $c = 1.1, 1.2, 1.3$, and the plot is shown below.

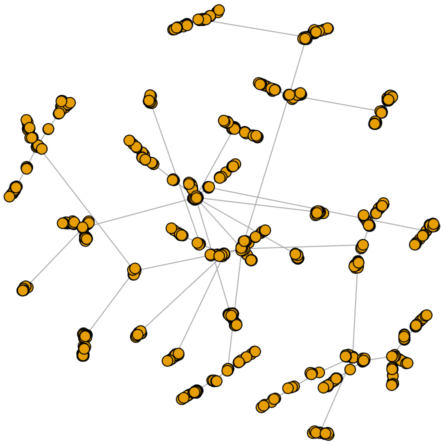


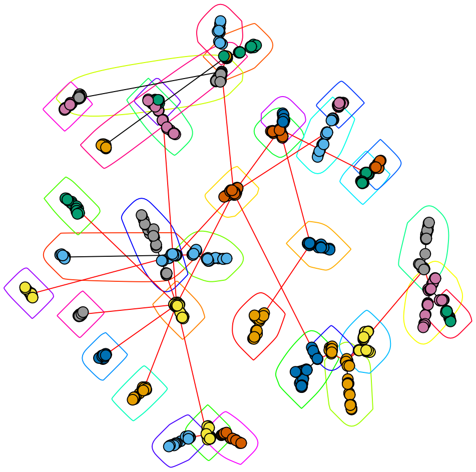
From the plot, we find that the expected size of GCC is almost positive proportional to the number of nodes n when $c = 1.1, 1.2, 1.3$, and the slope increases as c gets larger.

iv. The relation between the expected GCC size and n has been briefly revealed previous below each plot. To sum up, when $c < 1$, the relation is closer to a logarithmic curve; when $c = 1$, the relation between the two is in the transition from logarithmic curve to a linear line; when $c > 1$, it becomes a linear relation, and the slope increases as c increases.

2. Create networks using preferential attachment model

   (a) An undirected network with $n = 1000$ nodes using preferential attachment model is created as below, along with its community structure using fast greedy method, as required in question (b).

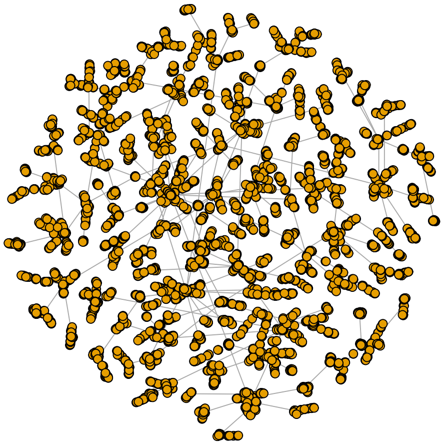


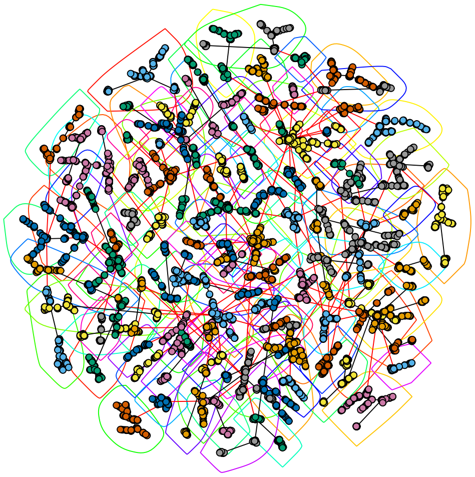


   According to the generation method in preferential attachment model, such a network is always connected.

   (b) The modularity of this community structure is 0.93361.

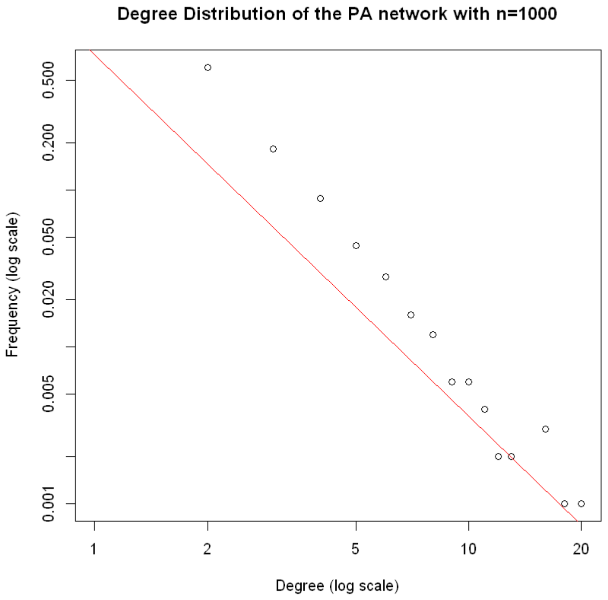
   (c) An undirected network with $n = 10000$ nodes using preferential attachment model is created as below, along with its community structure.

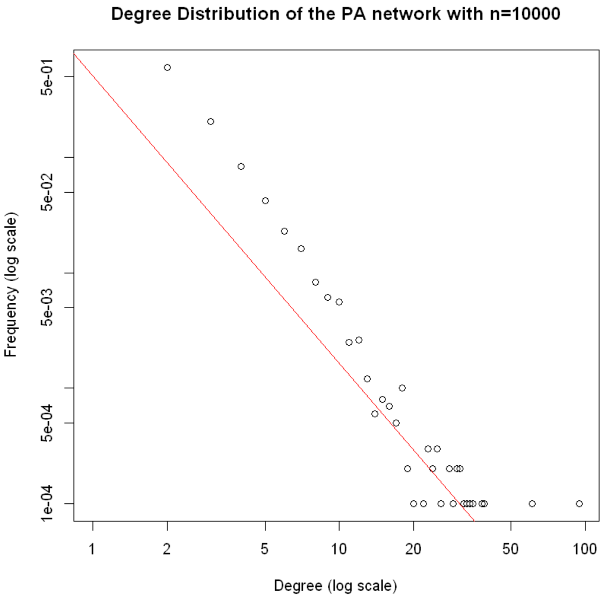




   The modularity of this community structure is 0.97817. It's larger compared to the smaller network's modularity.

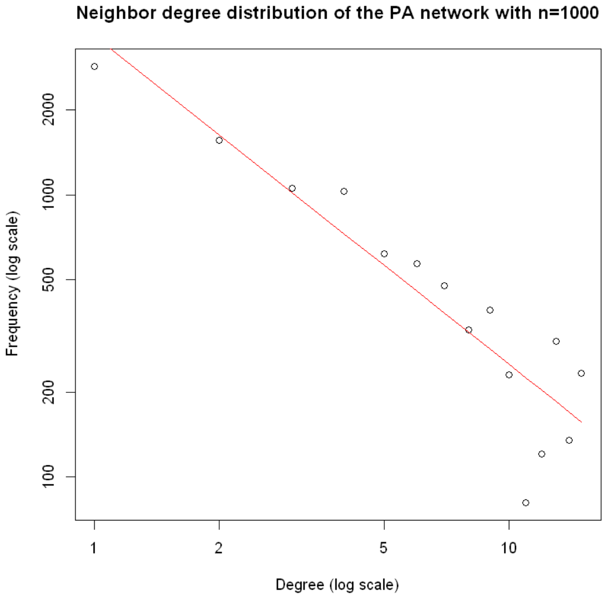
(d) The degree distribution for both networks with $n = 1000$ and 10000 are shown below.

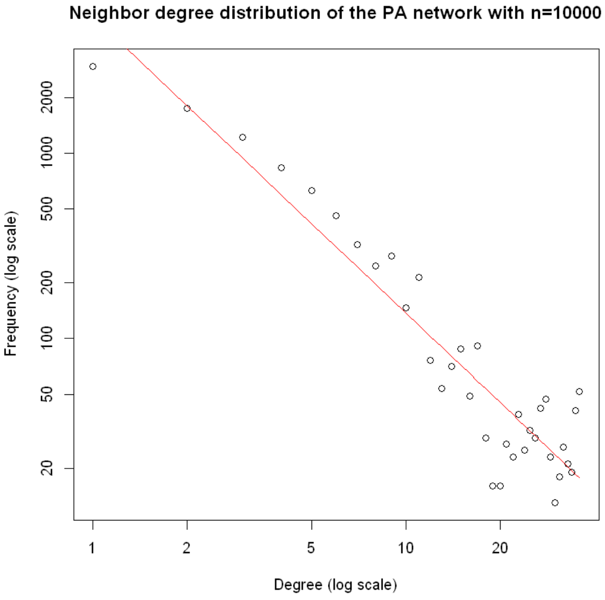

Degree Distribution of the PA network with n=1000


Degree Distribution of the PA network with n=10000

The slopes of the linear fitting lines are -2.1911 and -2.3598, respectively.

(e) The degree distribution of the one-step neighbor in both networks are shown below, with 10000 iterations in both cases.

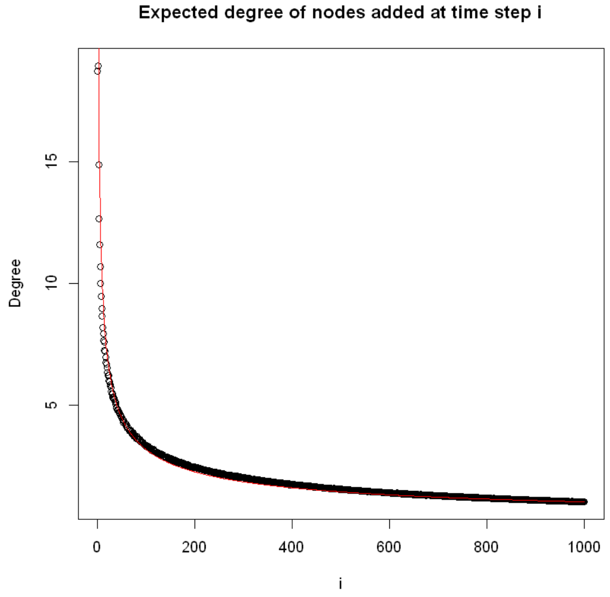

Neighbor degree distribution of the PA network with n=1000


Neighbor degree distribution of the PA network with n=10000

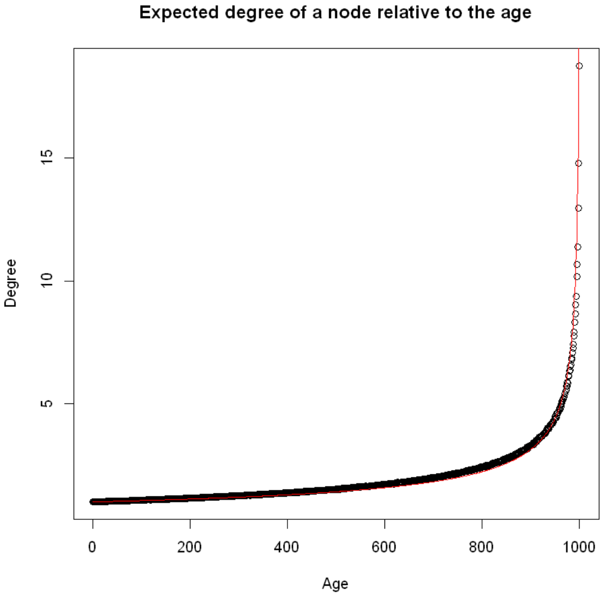
The distribution is linear in the log-log scale for both $n = 1000$ and 10000. The slopes are -1.149 and -1.611, respectively.

The slope from the neighbor degree distribution is smaller than the slope from the node degree distribution for both n. The fact that the slope from larger n is larger still remains in the neighbor degree distribution.

(f) We iterate 10000 times to estimate the expected degree value of nodes added at time step i for $1 \leqslant i \leqslant 1000$.



Expected degree of nodes added at time step i



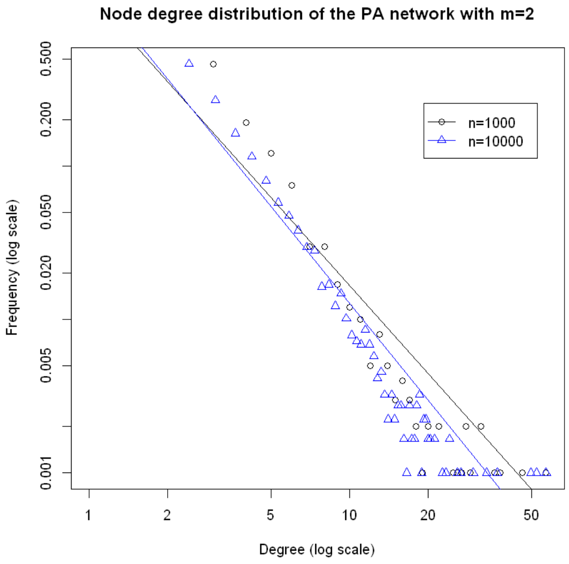
Expected degree of a node relative to the age

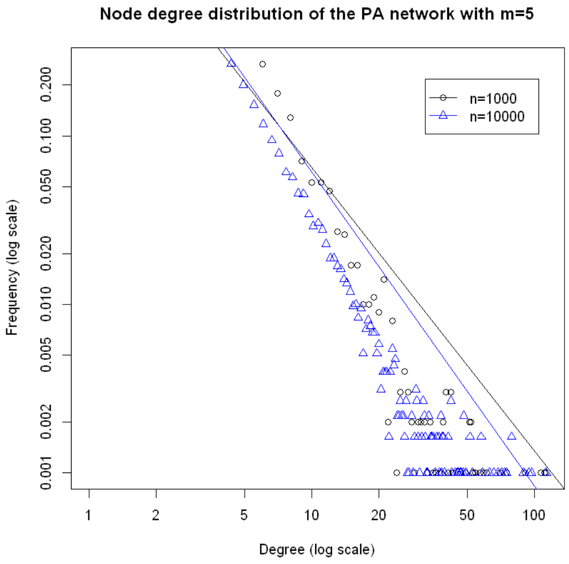
The "age" of a node is defined as the time step passed since the node is added, $\text{age}(i, t) = t - i = 1000 - i$. We plot the expected degree of a node added at time step i (on the left), and expected degree versus age (on the right), along with the theoretical fitting curve in red with the following theoretical equation:

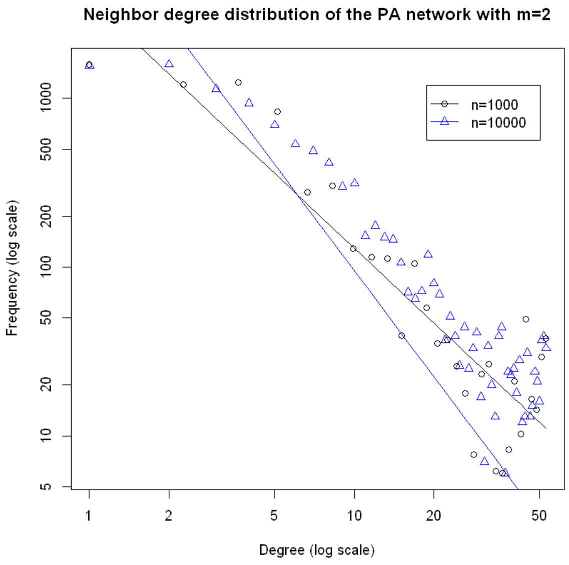
$$\text{Degree}(i, t) = m\left(\frac{t}{i}\right)^{1/2} = \sqrt{\frac{1000}{i}} = \sqrt{\frac{1000}{1000 - \text{Age}}}.$$

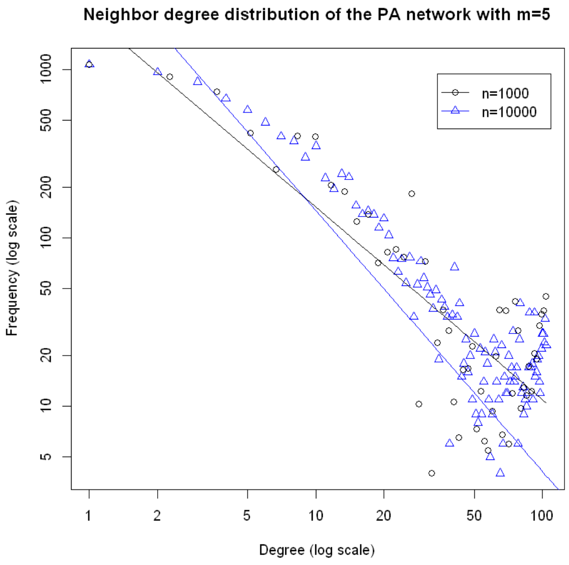
(g) We repeat the previous parts and create networks with $n = 1000$ and 10000 for $m = 2$ and 5, and here are the results and plots.

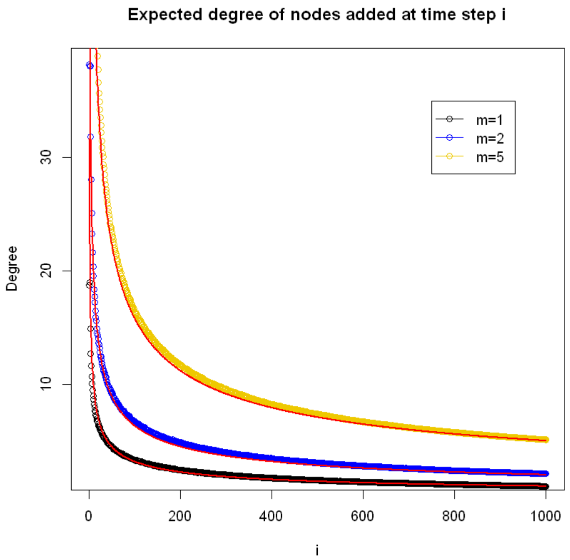
m	n	Modularity	Node deg. distri.	Neighbor deg. distri.
1	1000	0.93361	-2.1911	-1.149
	10000	0.97817	-2.3598	-1.611
2	1000	0.51808	-2.0917	-1.033
	10000	0.53161	-2.3050	-1.467
5	1000	0.27206	-2.001	-0.903
	10000	0.27543	-2.0913	-1.223


Node degree distribution of the PA network with m=2


Node degree distribution of the PA network with m=5


Neighbor degree distribution of the PA network with m=2


Neighbor degree distribution of the PA network with m=5


Expected degree of nodes added at time step i

We find that, for each fixed n, as m increases, the modularity decreases, the slope value of the node degree distribution and the neighbor degree distribution in log scale decreases as well. And the node degree distribution slope value remains larger than neighbor degree distribution for every m.

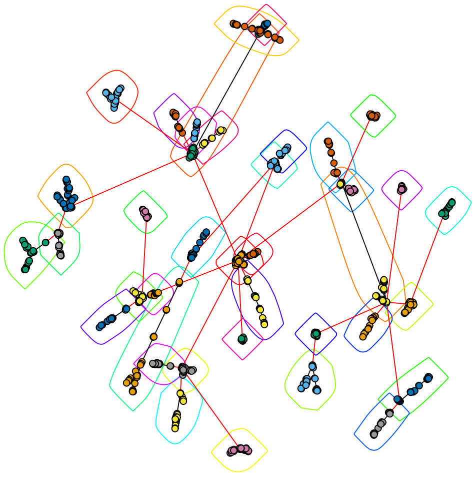
For each m, with larger n, the modularity is larger, the slope value of the node degree distribution and the neighbor degree distribution are larger, too.

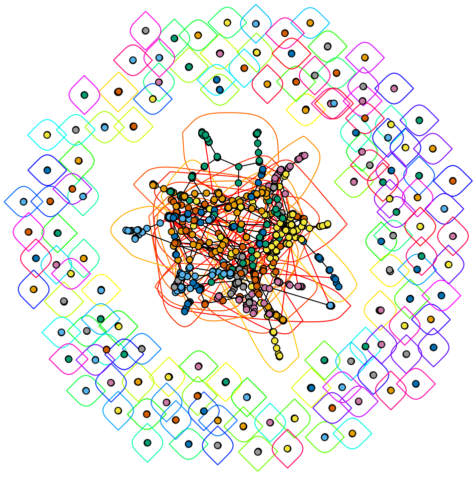
As for the expected degree of nodes added at time step i for different m, the curve always follows the equation:

$$\text{Degree}(i,t) = m\left(\frac{t}{i}\right)^{1/2} = m\sqrt{\frac{t}{t-\text{Age}}},$$

where $t = 1000$ in all three cases, and $m = 1, 2, 5$, respectively.

(h) We generate a PA network with $n = 1000, m = 1$ and a new network with the same degree sequence as follows.





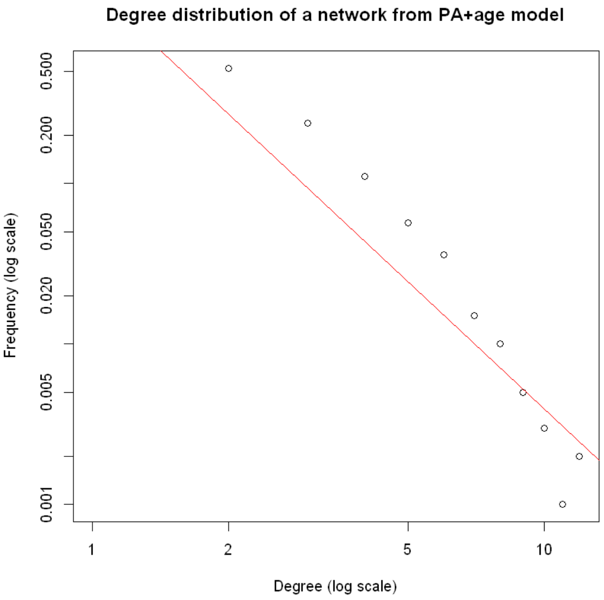
The modularity of these two networks are 0.92940 and 0.83502, respectively.

We find out that, the network generated from the preferential attachment model is ensured to be connected, while another network with the same degree sequence is highly likely not connected. From the plot we can see that there are many small components around the GCC of this network, therefore its modularity is smaller.

Even though the two networks have the same degree sequence, and thus both follows the rule of power-law, they have very different property, including connectedness and modularity.

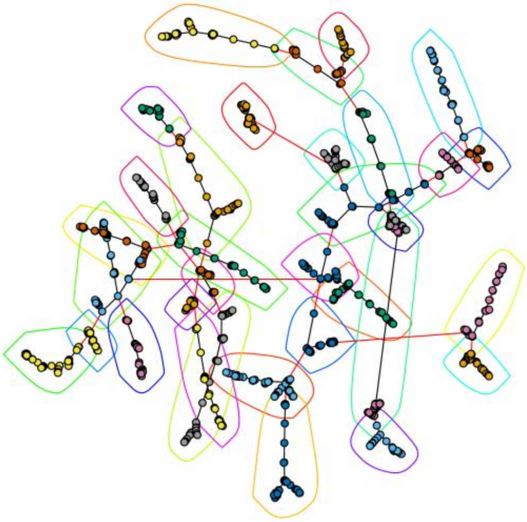
3. Create a modified preferential attachment model that penalizes the age of a node

   (a) The degree distribution of the network using modified PA model in log-log scale is shown below.



   The slope of the linear fitting line is -2.6561, which is also the power law exponent.

   (b) We plot the community structure of the network as follows.

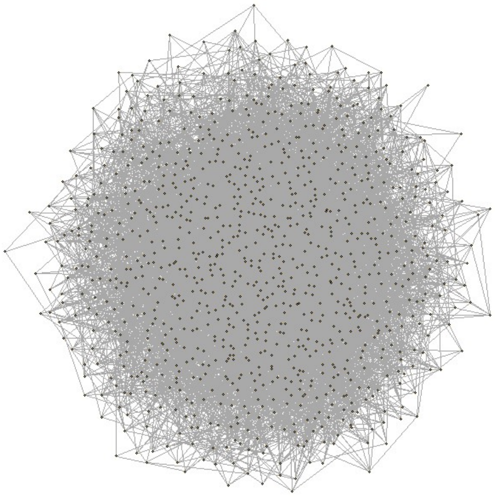


   The modularity is 0.93594.

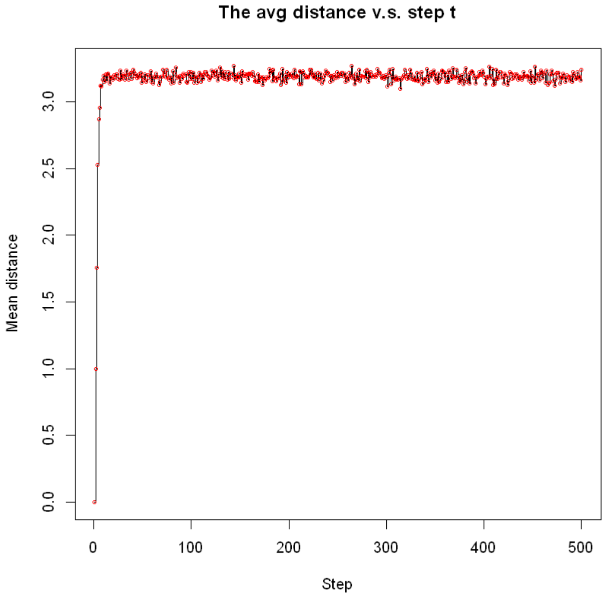
2 Random Walk on Networks

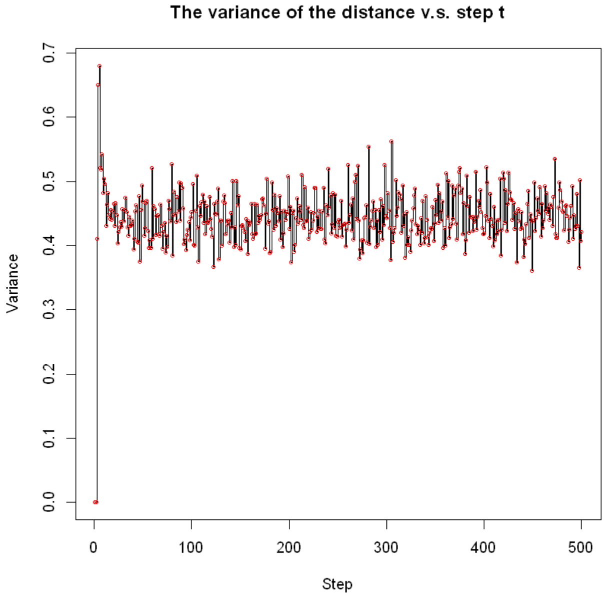
1. Random walk on Erdös-Rényi networks

   (a) We create an undirected random network with $n = 1000, p = 0.01$ using ER model.



   (b) Let a random walker start from a randomly selected node. For each destination, the shortest path lengths $s(t)$ are calculated. With different starting points, $s(t)$ for each t are different, and the mean distance and variance was measured over random choices of the starting nodes are measured. We use 500 steps for this 1000-node network, and iterate over 500 randomly picked starting nodes. The plots of mean and variance of $\langle s(t) \rangle$ are shown below.





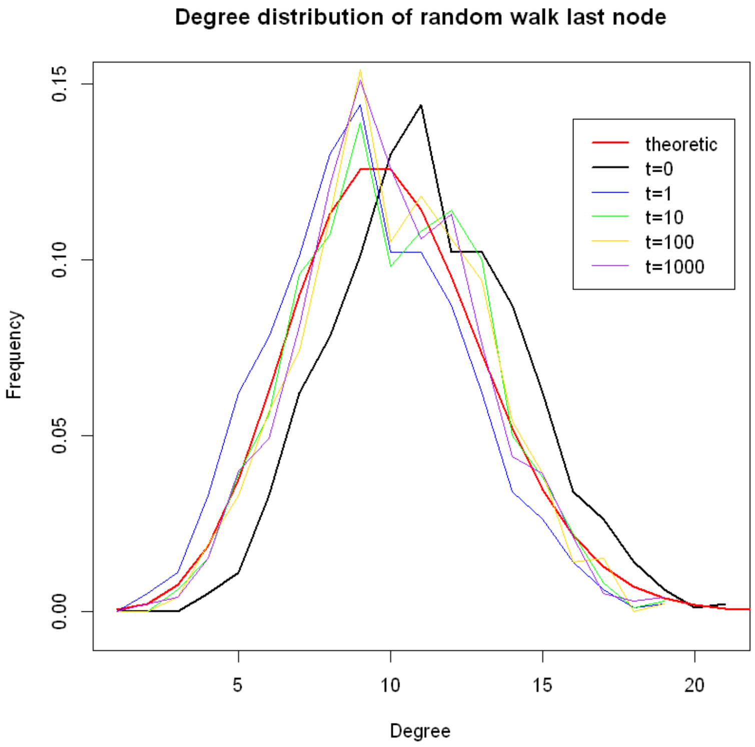
We find out that the mean distance of random walk is upper bounded by the diameter of the network $D = 5$, and both the mean value and the variance are stabilized after around 7 steps. The stable mean is about 3.188, the variance is about 0.444.

The theoretical average distance can be calculated from the formula below:

$$\ell = \frac{\ln(n)}{\ln\langle k\rangle},$$

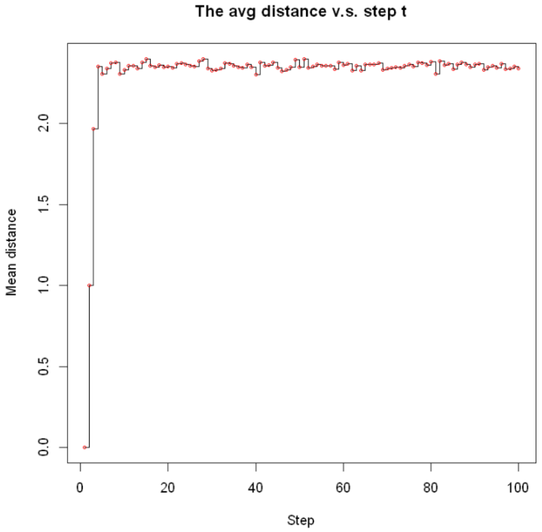
where n is the node number, $\langle k\rangle$ is the mean degree value, expected to be np. The theoretical average distance in our case is: $\ell = \ln(1000)/\ln(1000 * 0.01) = 3$. So our experimental mean distance from the generated network is close to the theoretical result.

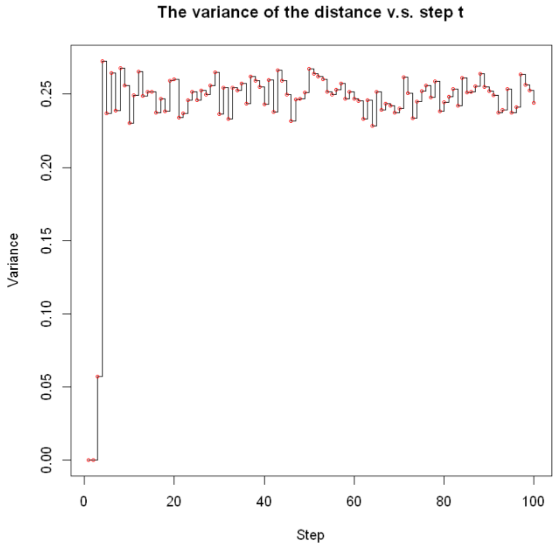
(c) We compare the degree distribution of the last node after step number of $t = 1, 10, 100, 1000$ to the original distribution and the theoretical binomial distribution, using 1000 random starting nodes.



By observation, there is not much difference between all degree distribution of the last node and the original network. They all approximately follow the binomial distribution with mean of $np = 1000 * 0.01 = 10$. The reason of the similarity is that the ER graph are generated with the same probability p for all connections between node pairs, therefore, each node can be considered identical in random walking.

(d) In this part, we increase the number of nodes from 1000 to 10000 and repeat the same process as in (b).



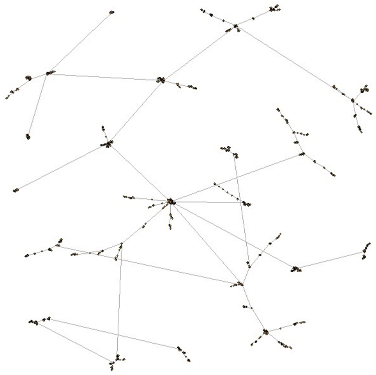


By observation, the $n = 10000$ network reaches the stable state with fewer number of step (around 4 steps), and both mean and variance fluctuates less as the step number t changes. The mean distance is also upper bounded by the diameter of the network ($D = 3$), and close to the theoretical average distance $\ell = \ln(10000)/\ln(10000 * 0.01) = 2$.

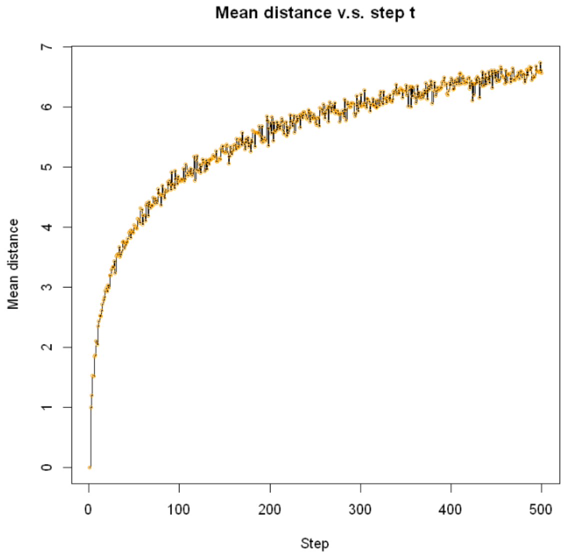
The reason why with larger node number n and the same connection probability p, the average random walk distance shorten, is that, the average degree rises, which means each node has more connections with other nodes. Qualitatively, since all the nodes are more connected, the diameter of the network gets shorter, by the definition of graph diameter, the distance between any two nodes is shorter than the diameter, therefore it is easier to find a shorter path from the starting node to the ending node of the random walk.

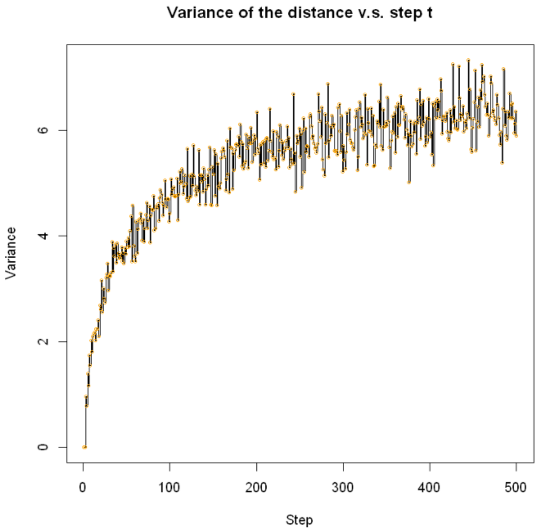
2. Random walk on networks with fat-tailed degree distribution

   (a) In this question, we generate undirected network using preferential attachment model with $n = 1000$ and $m = 1$.



   (b) Similarly, we repeat the steps mentioned in part 1, and plot the average distance and the variance of the distance as for the step number t on this new network



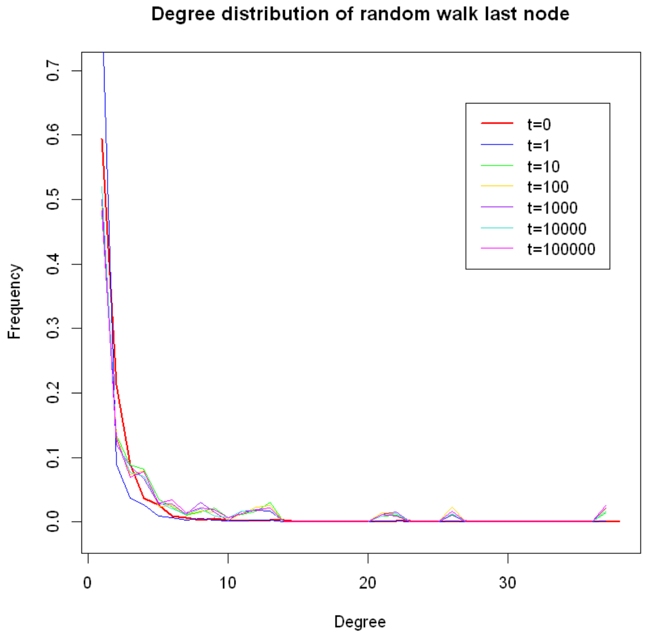


From the plot above, we can see that PA network's mean and variance reaches the steady state slower than ER's. The theoretical average distance can be calculated from the same formula:

$$\ell = \frac{\ln(n)}{\ln\langle k\rangle},$$

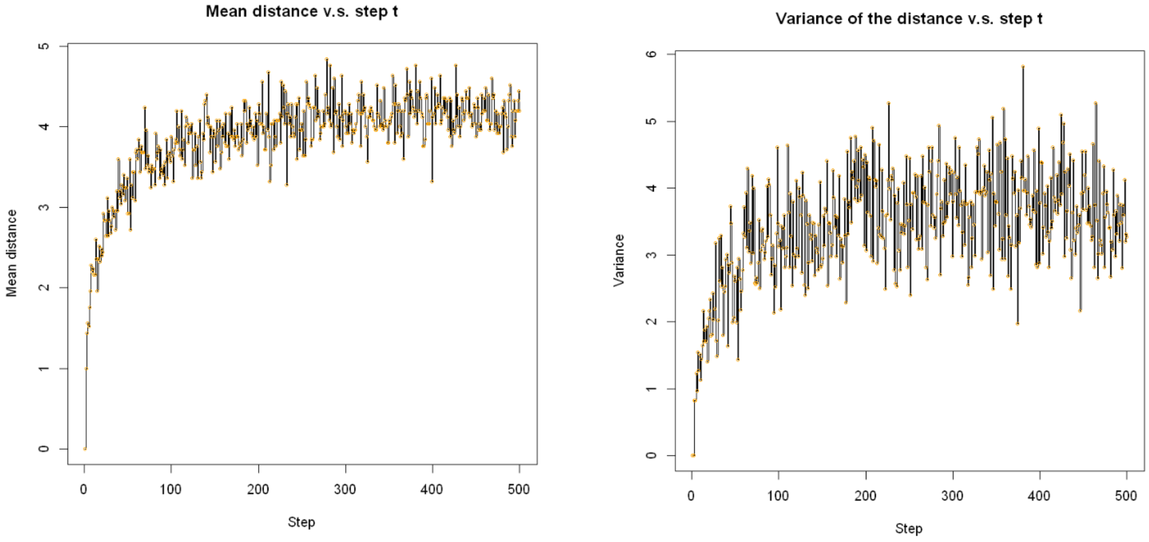
where n is the node number. The experimental value $\ell = 6$ is a little off from the calculated value. The variance also has greater value and more fluctuation than ER network.

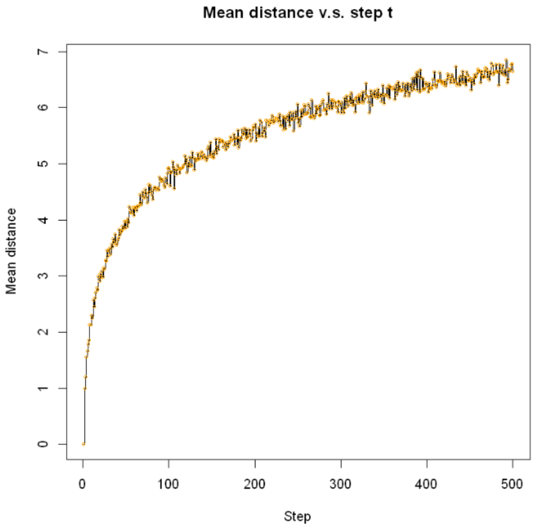
   (c) Similar to 2.1.(c), We need to calculate the degree distribution of the destination nodes and compare the degree distribution of the last node with that of the original graph. For PA network, the distributions converge much slower, so we chose to iterate over more choices of step number $t = 1, 10, 100, 1000, 10000, 100000$, and compare the degree distribution of the last node with that of the original graph.

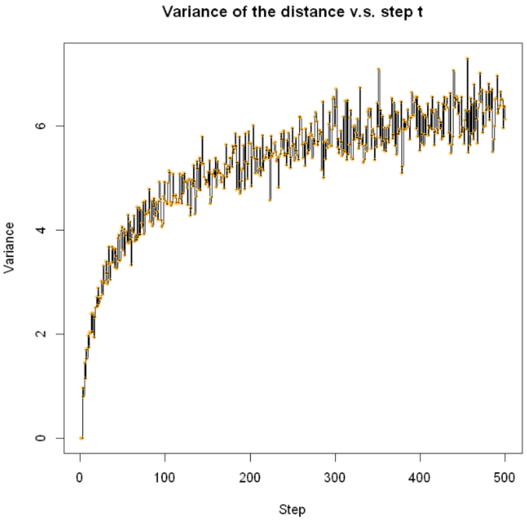


By observation, the degree distribution of all step numbers correspond to a power-law distribution, with different coefficients. The last node distribution converges to the distribution of the original network much slower comparing to ER network.

(d) In this part, we increase the number of nodes from 100 to 10000 and repeat the same process as in (b).



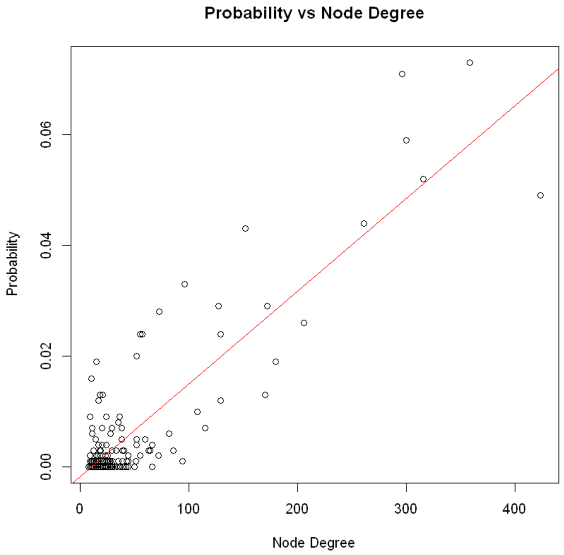




The average distance of $n = 100$ stabilizes faster than $n = 1000$ and $n = 10000$, and reaches a smaller value. Variances for $n = 100$ fluctuates more compared to variance of $n = 1000$ and 10000, which is just because it does not have as many starting nodes to calculate the average. Similar to ER networks, the node number, diameter and mean distance are positively related for PA networks.

3. PageRank

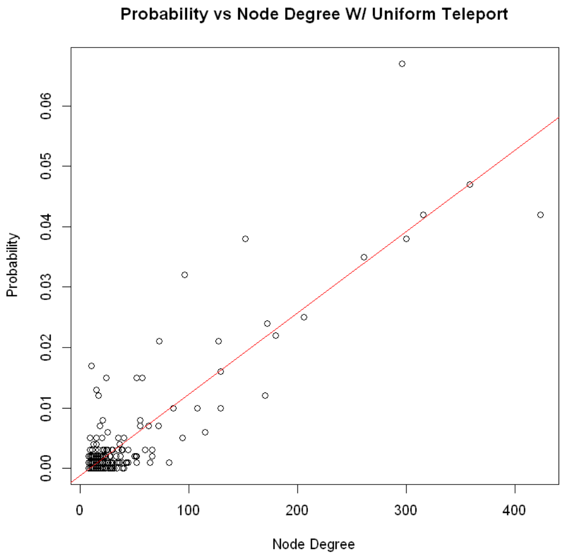
(a) For the page rank algorithm, we constructed two directed networks using the preferential attachment model and merged the two networks as instructed. This is done in order to avoid a black hole in the random walk process since the first node of a directed network has no outbounding edges. After shuffling and merging the edges, a transition matrix A is constructed and normalized such that each row represents the probability of walking to node i from every other node j. Then, a random walk is performed by sampling from the probability distribution of the outgoing edges. Each end node after 1000 steps are recorded and 1000 random walks are iterated through to generate the following figure of probability versus node degree graph.



The figure above shows that the nodes with a higher degree are more likely to get visited and have a higher pagerank score.

(b) To consider teleportation, the probability distribution of the random walk process now contains two parts: the original random walk probability distribution and an uniformly distributed teleport probability for all nodes. The teleportation probability is scaled with a teleportation factor, α, which is set to 0.15 in this study, and thus a scale of $1 - \alpha$ is set to the original standard random walk probability distribution. This means that there is a 15% chance that the walker will teleport to any node in the network, and all the nodes (except for the current starting node) in the network have the same probability to become the next node.

Using this new model, we plot the relation between the probability of a node to become the ending node after 1000 steps and its degree as follows.

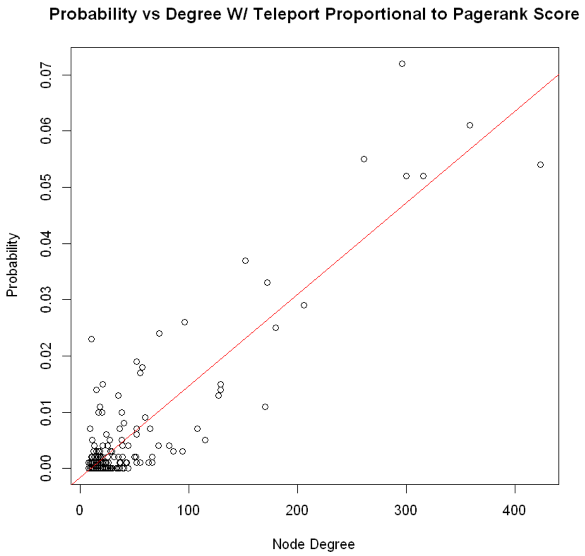


With an uniformly distributed teleportation probability, nodes with small degrees now have a higher probability of getting visited. On the other hand, quantitatively, the slope of the linear fitting curve between probability and node degree in 3(a) is $1.671 * 10^{-4}$, and it is $1.343 * 10^{-4}$ after using uniform teleportation, suggesting that the nodes with high degrees now have a lower probability of getting visited.

4. Personalized PageRank

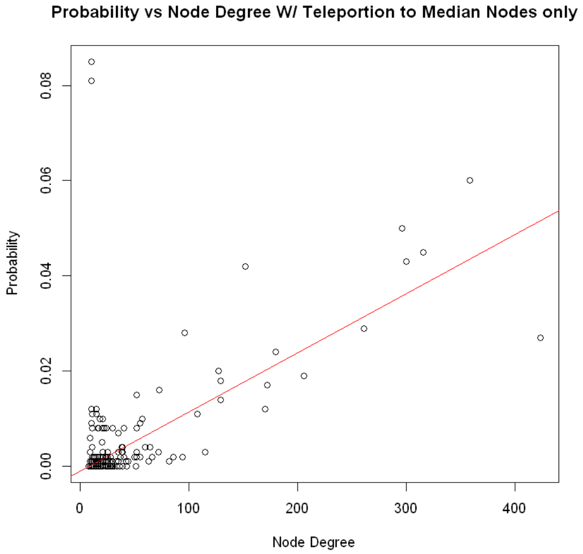
   (a) Now in this part, we further modify the teleport matrix in the way that the teleportation probability is proportional to the pagerank of each node, and the chance of teleportation versus standard random walk remains 15% and 85%.

   Using this new model, we plot the probability of arriving at the node versus its degree as follows.



We find out that the linear fitting line is of a similar shape compared to 3(a). It is easy to understand because the algorithm assumes the random walker has a higher probability of teleporting to nodes with higher pagerank, but these nodes don't necessarily have higher or lower node degree. Therefore, the probability doesn't have an obvious change with regard to node degree, compared to the situation without teleportation.

(b) In this model, we first find out the two nodes with median page rank, and then design the teleportation matrix so that the random walker could only teleport to these two nodes. The resulting plot, along with the summary table of linear fitting coefficients in the above four models, are shown below.



	slope($*10^{-4}$)	R^2
no Teleport	1.671	0.7939
uniform Teleport	1.343	0.7971
Teleport ∝ PageRank	1.626	0.8172
Teleport to Medians	1.241	0.4275

We notice that, by limiting the teleportation only to the two median-pagerank nodes, the probability-degree plot has more scattered points. There exists some nodes with low node degree and high visiting probability, and also nodes with high node degree and comparatively low visiting probability. Quantitatively, this model has the lowest slope and R^2 for the linear fitting curve. The reason is that, in this model, the random walker would frequently teleport to the two designated median nodes, and therefore nodes close to these two nodes would have a higher probability to be visited.

(c) The original PageRank equation is:

$$\pi P = \lambda \pi,$$

where P is the transition matrix, and π is the steady-state probability matrix of the random walk with teleporting and thus the PageRank value, λ is the eigenvalue.

To take the effect of self-reinforcement into consideration, we think a new parameter indicating the strength of this self-reinforcement γ can be introduced, and the modified equation would be"

$$\pi^{\gamma} P = \lambda \pi^{\gamma}.$$

The more the random walker self-reinforce, the less possible it is to explore new nodes, and the PageRank would appear higher and higher for certain familiar nodes and decreases exponentially for other nodes, just like how the exponential part normally works.